Sandeep Avula. Which Hashtag to use? Building a Hashtag recommender system and understanding the textual features surrounding Hashtags. A Master's Paper for the M.S. in I.S degree. April, 2014. 37 pages. Advisor: Jaime Arguello

Hashtags are community-based tags on twitter that are used to annotate tweets and make them findable. To make a user's participation on the social platform more relevant, recommending a hashtag would help a user participate better. This study is an attempt to build a recommender system for hashtag recommendation, and to further study the textual features around hashtags, which assist in their retrieval. The suggested system performs better for tweets with longer text; those with a URL, with multiple hashtags and those that have user mentions.

Headings:

- Microblogs
- Folksonomies
- Query Likelihood
- Mean Reciprocal Rank
- Priors
- Indexing
- Recommender system
- Search Engines

WHICH HASHTAG TO USE? BUILDING A HASHTAG RECOMMENDER SYSTEM AND UNDERSTANDING THE TEXTUAL FEATURES SURROUNDING HASHTAGS

by
Sandeep Avula

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2014

Approved by

_____

Jaime Arguello

Table Of Contents

Introduction

With the continually increasing number of conversations in the social sphere of micro blogging platform twitter, it is a great challenge to give power to the users to follow conversations. To keep up with the massive amounts of conversations, hashtags have emerged as an ontology that can facilitate users. Hashtags are essentially words prefixed with a '#', which people can insert in their tweet (Tsur, O., & Rappoport, A., 2012). For example to indicate that content is funny one can use #LOL or #Funny or the many such variants, and to report something sad one could possibly user #SAD or #hate. Given that tweets consists of only 140 characters these hashtags have intuitively become text that add more context than what the word used actually means.

Using a hashtag to label a topic or conversation isn't novel to twitter. It was extremely popular in Internet Relay Chat (IRC) networks (MacArthur, A). Hashtags in the context of the IRC networks too provided depth and context to conversations. Twitter adopted the use of hashtags officially in July 2009. This adoption took place after people familiar with the IRC style usage of hashtags began using it on twitter.

Since twitter has adopted the usage of hashtags, people have adopted this new syntax for conversation with many using them even in regular conversations. Even services like

Instagram (Van Grove, J., 2011), which majorly convey the subject via pictures, use hashtags to add context to the picture.

Tags

To comprehensively understand the usage of hashtags, tagging as a practice itself needs to be understood. Tags typically were used to categorize text, and help indexing by adding metadata that was directly curated by the users. Adding manual metadata is often costly (Di Caro, L., et al, 2008) and with people becoming more active on the Internet it only made sense to let them add the metadata they thought would help them better identify articles they read or web pages they visited. But the tags that people could use were controlled (Quintarelli, E. 2005) and given the rate at which people became active on the internet, these rigid hierarchies could not keep up with the needs of people – this was especially evident with the spawning of several social media platforms like Delicious and Flickr (Gupta, M., Li, R., et al, 2010). To heed to the needs of the users new concepts like social tagging and folksonomies got introduced. Social tagging was a methodology that gained prominence after the launch of the websites such as, Flickr and Delicious (Marlow, C., et al, 2006). It essentially gave the users the ability to tag items with names they felt were appropriate, and based on those names, similarly tagged items could be searched. Folksonomy was a word coined by Thomas Vander Wal in the AIfIA mailing list (Gupta, M., Li, R., et al, 2010), where the concept is based on the idea of a flat name space. It alienates the need to explicitly define relationships between terms, and the users had the freedom to choose the names they felt appropriate. This new way of tagging was fundamentally different from the previous methods in the sense that the previous method

of tagging took a top down approach, where hierarchies were already built and people used the given structures build downwards; and in the new system, it is bottom up.

Hashtags on twitter are based on the combination of social tagging and folksonomies. A crucial difference is that hashtags are used right within the text, unlike in the previous usage models where they essentially were more like headings of sorts, or added as labels separate from the actual content. Moving on to a larger question, are the hashtags used to just give conversations more context? As it turns out they don't, and as pointed out by (Yang, L., et al, 2012), hashtags serve a dual purpose of marking content and as a symbol of inclusion in a community. Hashtags basically define a virtual user community of users with either similar interests, or backgrounds or basically anything that can further and enhance the community. For example if someone would use a hashtag "#sigir2013", it could be said with certain confidence that the people using this hashtag are referring to the SIGIR 2013 conference and belong to academia or have interest in it. Essentially by using a hashtag a user can expect to not only mark their conversation but also put forth their intent to participate in a community. There are extreme cases where hashtags have been used for more serious purposes, for example during government elections and natural disasters (Potts, L., et al, 2011). By participating under certain hashtags, people have been able to share information with interested folks across countries, time zones and cultures. Journalists too use twitter as a medium to quickly connect with their audience about quick newsworthy events, and adding a hashtag to their content makes it easier for people for search for the information (Farhi, P., 2009). For example during the Boston Marathon bombings, the hashtag #bostonmarathon quickly changed from being a hashtag

used to encourage and congratulate participants to being a marker related to the bombings. Essentially people intuitively leveraged upon the fact that there was a community built around that particular hashtag and went ahead to warn people via that hashtag (Spero, S., 2013).

Given the context above, the next question I shall delve into is –Given a particular situation how can one figure out the apt hashtag to use? And to build such a system, what properties of a tweet facilitate hashtag recommendation? And coming to asking a critical question – Would a hashtag recommender system assist people? Previous work by (Sen, S., et al, 2006) has shown that people do use suggested tags. The information foraging work by (Pirolli, P. 2005) too has suggested that people tend to adopt suggested tags to optimize the information/effort ratio, meaning that they would optimally put the least effort to get the best possible information. And (Sen, S., et al, 2006) too have shown that a communities tagging behavior greatly influences users tagging behavior, so suggesting them the widely used hashtags should provide users a better online experience.

Related Work

Suggesting tags isn't a new field of study; there have been studies pertaining to suggesting tags for webpages, and prior to twitter, on social media platforms like Flickr and Delicious. A central thing that needs to be understood in the perspective of twitter is the tweet size. It is interesting to see how researchers with a larger body of context prior to micro blogs have worked on recommendation systems and how after the introduction of the microblog research has been driving to tweak or bring in new techniques to build recommender systems.

Prior to the emergence of social platforms like Delicious, Flickr and Twitter, traditional blogs were a great interest of study. The approaches all come under three primary categories: Content Based approach, Collaborative approach and a Hybrid of Content Based and Collaborative based. Collaborative approach is one that has gained more prominence with the growing access to data pertaining to different users. One work that was based on collaborative filtering is AutoTag (Mishne, G. (2006)), where tags were suggested for blog posts. After a blog post has been written similar blog posts written by different users was searched for, and the tags used by them were aggregated. Amongst the aggregated tags, their frequency of usage was taken to give a better score. For example, if three people used a particular tag X, and 2 people used a particular tag Y. X would be scored over Y. Further based on the users own usage of tags, those tags the

user used in the past were taken into consideration to give a boost by a constant factor if those were to found in the aggregated list.

In the previous work we see that tag ranking gave preference to the frequency of usage of tags. There has been another work FolkRank (Hotho, A., et al (2006,)) applied for tags of bookmarking application Delicious, where the central idea is that a resource being tagged by an important tag by an important resource becomes important. PageRank has inspired this algorithm, where the central idea is that a webpage is important if more pages linked to it, and if those pages are important themselves. The importance is basically derived after multiple iterations, and not something that is available at the very beginning. A key take away from this approach is the drive to move away from a vector space model that is dependent on TF/IDF weights, as short snippets of text do not provide a great resource to compute TF/IDF's. This is a useful angle as tweets too are small snippets and as will be seen later, where a likelihood model has been as the basis for the recommender system. Another interesting area of research has been suggesting tags for a movie that offers users a brief look into what the movie might actually be (Sen, S., et al (2009)). A users tagging behavior in the past mostly power this system.

Now coming to our main area of interest, twitter. There have been different types of strategies ranging from likelihood models, query expansion and topic models to make the best possible recommendations.

(Efron, M. 2010) approached the task as a form of entity search. Initial ranking of the hashtags was done by taking a tweet as a query and measuring its KL-Divergence from the model of each hashtag, that was constructed by the tweets associated with that hashtag. Of the top 25 suggested hashtags query expansion was done, where the top 25 suggested hashtags were collected and a vector of them was constructed. The expanded query was constructed by linearly associating the initial query model for the tweet with this new model constructed by the hashtag. The parameter used in the linear association was empirically tuned. For the model generated from the recommended hashtags by the KL-Divergence two models were generated where in one, all the weights of the vectors were of equal weights, or assumed them all to be of similar distribution, and in the other model the weights were taken to be proportional to their IDF ratio (the hashtag word's) by the max IDF ratio. The data for this study was obtained using Twitter's API over a 24-hour period and 29 topical tweets were created based on the author's interaction with Twitter. For judgments Amazon's Mechanical Turk was used there each hashtag was to be rate on a 0 to 3 scale (not useful to most useful). 5 users judged each query-tag and the relevance were graded by NDCG. Based on the results, the feedback model performed statistically better than the initial model.

In a study by (Kiwi, S. M., et al, 2012), the user preference along with the tweet content was taken in account to make hashtag recommendations. The corpus used was built by analyzing tweets written by over 150,000 Singapore users over a three-month period from October 2011 to December 2011. A user profile vector was constructed by representing a user by all the unique hashtags captured. And the weights to these vectors

were based on the users affinity to use a hashtag. And a tweet too was similarly build by representing tweets in terms of a vector of all the words. Tweets similar to the subject tweet were taken and the hashtags used in those tweets were considered. These were grouped along with the hashtags used by users who were similar to the users profile. A peculiarity in their approach is the use of TF/IDFs where the frequencies were based on a users usage of a hashtag. For example, the IDF is a log ratio of the Total number of users, to users who have actually used that particular hashtag. To evaluate their system performance, they used what they called a Hit Rate, which was the ratio of the number of actual hits (their systems predictions) to actual total number of target user-tweet pairs.

In the paper by (Zangerle, E., et al, 2011), for a given tweet they have searched for tweets that are similar to the given tweets. Out of the most similar tweets, the hashtags were extracted and these hashtags were listed out as recommendations. The similarity between tweets was calculated by using a basic cosine similarity, where the tweets were weighted by using their TF/IDF's. Of the recommended hashtags, these were ranked in three ways. One, was based on the hashtags overall popularity over the entire corpus, two, was based on the popularity of the hashtags of the recommended hashtags, and the third was based on the similarity of the subject tweet, to the tweet that produced the hashtag. The evaluations were done based on precision and recall. Precision here means that a depth of K, how many of the original hashtags used in the tweet were recommended. I have my reservations to this methodology as at a depth of greater than 5 (they have taken K from 1 to 10), if we find the right match, the precision would be 1, but that doesn't tell us how

good the system is. Perhaps a better approach could have been to take the Mean Reciprocal Rank (MRR).

Another interesting approach to this task was taken by (Ding, Z., et al, 2012), where they used topic models and translation models to make the suggestions. Their efforts centered around finding the right hashtags for a tweet that came under similar topic and whose previous usage was close to the topic-specific word alignment table. Another somewhat similar approach was taken by (Godin, F., et al, 2013), where they've applied an LDA model to make the hashtags recommendations. They used human evaluators to make the evaluations for their judgments.

Problem Statement

The goal of this study is to build a recommendation system and understand the impact of features related to the construction of a tweet such as tokens (words, numbers and symbols), URLs, hashtags and user mentions. For each of these entities there is a motivation on why they are being considered. A central assumption here is that there needs to be a minimum of one token in the tweet for the system to suggest something.

Tokens: To build a context or a topic model around a hashtag, one typically needs tokens (words, numbers and symbols) that are being used along with the hashtag. These tokens essentially form the core of the content. So to start this study, the first basic step is to analyze if tweets having tokens more than 9 perform better than those having tokens less than that.

Hashtags: Hashtags are either used singularly inside a tweet or along with more hashtags. Typically this is done to add more contexts to the tweet. So, essentially what we shall drive from this is to check if it becomes easier for the recommender system to predict hashtags for those using multiple hashtags or those using singular. This is helpful as this in the longer run indicates that people can substitute content in the form of tokens to hashtags.

URL's: Twitter allows users to embed URLs in a tweet so that they may share content on the web with the other users. Now what would be interesting to see if having a URL in a tweet makes it easier for a system to make a recommendation. This then takes us to answer a further question – Why does this work/not work?

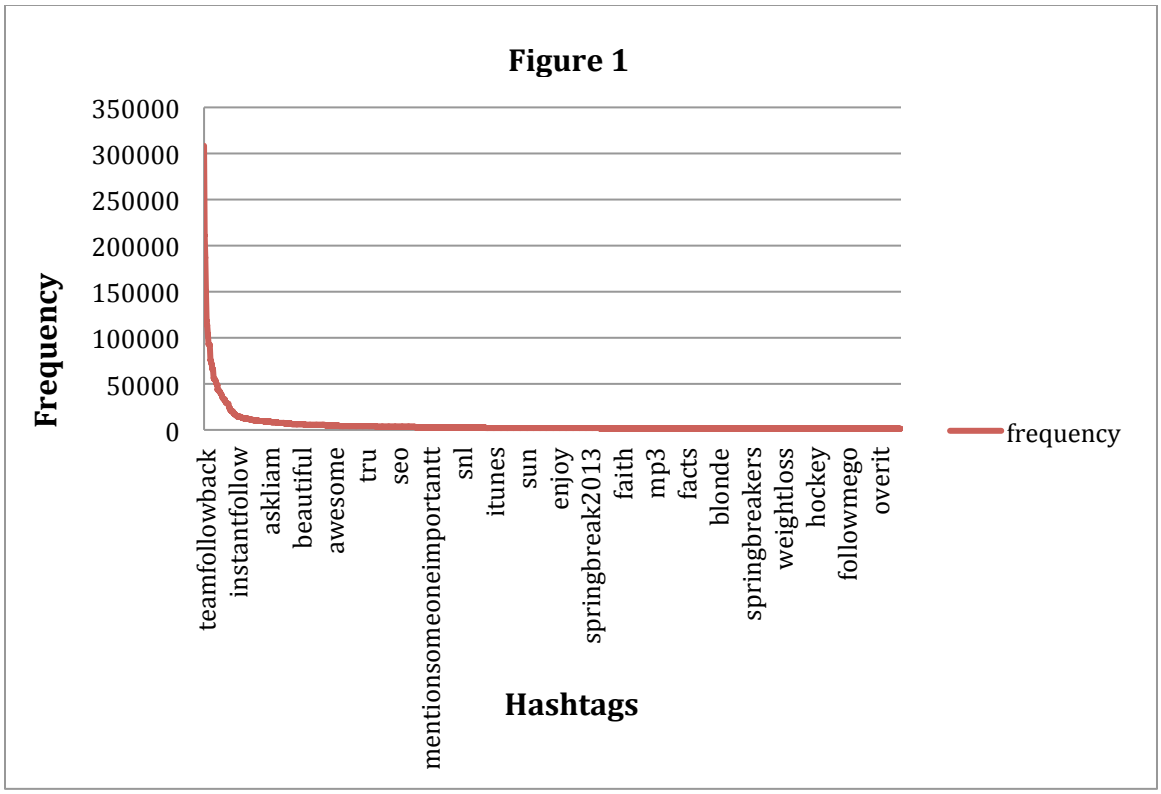
User Mention: A user mention in the twitter vocabulary means that another user gets notified that the tweet that has been sent out as one directed towards that person. For example a tweet –"Hey @user_123, how are you doing?" is a tweet that is directed towards a person with the twitter handle "user_123". And we shall try to understand if tweets directed towards users perform better at recommending hashtags and understand possible reasons of why it works/doesn't work.

The System: To work on the scenarios above, there would be two systems developed. One a baseline system and another that would apply a prior based on the hashtag usage distribution.

Data Set

The data used for the experiments are tweets that have been collected in the early part of 2013. The tweets used for these experiments are between the dates 03-01-2013 to 04-01-2013. The tweets were gathered using twitter's standard API. Candidate hashtags were generated from the tweets on 03-31-2013. For each candidate hashtag, a language model was constructed using all tweets containing that hashtag published in the past 30 days (03-01-2013 to 03-30-2013). Only English text was taken into consideration during the construction of the language models and the total number of hashtags collected is 108,784.

A peculiar observation from this data set has been that of the 108,784 candidate hashtags, it was not possible to construct a language model for about 42,237 of them because they did not occur in any tweets published in the past 30 days. Below is a graph showing the distribution of the frequency of usage of the hashtags being analyzed. As it can be seen, the distribution is very skewed and follows the power law.



Test set

To satisfy the needs of the problem statement, test sets have been made. First a central test set containing all the tweets to be used for testing has been made. That file has then been split as per the problem statement and has been described below.

Tokens: To work on the analysis of the length of the tokens used in a tweet, 2 files were created, where one contains tweets having tokens less than or equal to 9 and the other having tokens greater than 9.

Hashtags: Two files have been created with one containing only those tweets that have a single hashtag, and with the other containing multiple hashtags.

URL's: Two files have been created with one file containing tweets that have a URL and another containing those tweets which do not have a URL.

User Mentions: Two files have been created with one containing those tweets that have a minimum of one user mentioned, and another with no mention of a user.

Test File	**No. Of Tweets**
Main Test File	18,993
Test File with Multiple Hashtags	5,229
Test File with Single Hashtags	13,764
Test File with tokens greater than 9	10,130
Test File with tokens less than or equal to 9	8,863
Test File that with no user mentions	8,539
Test File with user mentions	10,460
Test File with URLs	3,515
Test File without URLs	15,469

Algorithms

There will be two systems called Baseline System and Prior System.

Variables used:

$H = \{h_1, h_2, \ldots..h_n\}$, where H is the set of all hashtags containing n hashtags.

$T = \{t_1, t_2, \ldots.t_m\}$, where T is the set of m Tweets taken in consideration for the experiments

$t1 = \{t_{11}, t_{12}, ..t_{1j}..t_{1k}\}$, where t1 is the first tweet, and t_{1j} is the j^{th}, word in t_1.

Baseline System

For this baseline system I shall use a standard query-likelihood model that essentially works like this:

$$Score_A(t_i, h_j) = \prod_{k=1}^{m} (P(t_{ik}|h_j))$$

t_{ik} is the k^{th} term in the tweet t_i, and h_j is the j^{th} hashtag.

Smoothing

The algorithm above had no smoothing, and Dirichlet Smoothing has been employed. So the probability below has been used in place of the one described in $Score_A$

$$P_\mu(t_{ik}|h_j) = \frac{(tf_{t_{ik}} + \mu P(t_{ik}|h_j))}{(|d| + \mu)}$$

$tf_{t_{ik}} =$ Term frequency of t_{ik}

$|d| =$ Total words in the document.

$\mu =$ 2000, this value has been set after some empirical tests.

The above smoothing technique has been describe by (Zhai, C., & Lafferty, J. 2001), and would help us get a better estimate than by using the query-likelihood equation mentioned above as it is.

Prior System

This system is represented as follows:

$$Score_B(t_i, h_j) = Score_A(t_i, h_j) * Prior(h_j)$$

$$Prior(h_j) = \frac{f_j}{\sum_j^n f_j}$$

This prior basically gives information pertaining to the usage of a particular candidate hashtag amongst all the candidates. But this prior cannot be used as shown above, because of the skewed distribution shown in Figure-1. Instead, the logarithm of the value has been taken; as a log of the prior distribution produces more linear continuous values that can be taken advantage of. Hence, the new prior is as follows:

$$Prior(h_j) = \frac{\log_{10} f_j}{\sum_j^n \log_{10} f_j}$$

As the denominator is a constant, we will settle with calculating only the numerator.

Evaluation Metrics

From the previous studies (Efron, M. 2010) (Godin, F., et al, 2013) (Ding, Z., et al, 2012), we see that they've either taken the KL-Divergence or have put the system before users to conduct a user study on the performance of their proposed systems. When taking user studies, they've used Precision and Recall as their primary performance metrics. I believe that given the randomness and the veracity of the audience that use twitter, a better means to understand the performance of the system would be to try to predict the true hashtag that was actually used. So, keeping that in mind the Mean Reciprocal Rank (MRR) has been used to evaluate the system. For the given task, this is a harsh methodology to rank as this metric essentially gives information pertaining to the rank at which we can get the true hashtag, but nothing else pertaining to the information surrounding the hashtags suggested in the proximity of the true hashtag. So, what we are evaluating is not the number of useful recommendations, but the ability of the system to predict the true hashtags.

Experimental Setup

The code was primarily run on the KillDevil cluster at UNC-Chapel Hill's computing services. Indexing on the documents was done using Lucene 4. The test sets mentioned above were run against two algorithms, one was the baseline and the other with the prior.

At first the hashtags were collected and files were made based on the names of those hashtags. The text pertaining to each of those hashtags as mentioned earlier was then appended into those files. This content of each file was the language model pertaining to the respective hashtag. And as mentioned earlier, only English text was considered. Indexing was then done on each of these files, and then Lucene's inbuilt scoring libraries were utilized to perform the scoring.

To calculate the MRR of each tweet, the tweets were run against the index as per the algorithm and all the hashtags were collected. The rank of the relevant hashtag was then marked and the respective MRR calculated. To calculate the average MRR, all the MRR's pertaining to the tweets in the test set were aggregated and the average was taken.

Adjusting Complexity and approximations

Ideally while evaluating the MRR, it is required to make an exact match of the hashtag. For example, if the correct hashtag was at rank 10,000, the MRR would be 1/10,000, but

statistically that is not very useful as the number is too small. Added along with the statistical insignificance would be the unnecessary time and space complexity of the algorithm. To make the algorithm faster and to lie within statistical bounds, restrictions have been placed, where the last rank to be analyzed for the MRR would be 500, and anything below that would be directly ranked as 108,784, which is the total number of hashtag candidates.

Results

As mentioned earlier the evaluation metric for this study is the Mean Reciprocal Rank (MRR), and for a given tweet, the best possible MRR is 1 i.e., the correct hashtag is the very first one that is predicted. If the correct hashtag would be the second recommendation the MRR would be 0.5, and as per the experimental settings described above, the worst MRR would be 9.192E-6, which is almost zero. For each test scenario, if a suitable hashtag was found within the first 500 results, the MRR value was calculated; otherwise the rank was set to a default of 108,784 to penalize the bad performance and to improve the time and space complexity of the algorithm.

Test File	**Baseline System**	**Prior System**
Main Test File	0.279	0.311
Test File with Multiple Hashtags	0.438	0.505
Test File with Single Hashtags	0.218	0.237
Test File without URLs	0.237	0.267
Test File with URLs	0.461	0.496
Test File with user mentions	0.302	0.325
Test File without user mentions	0.250	0.292
Test File with tokens less than or equal to 9	0.228	0.271
Test File with tokens greater than 9	0.321	0.344

Below are graphs that show the distribution of ranks for the Baseline System and the Prior System. A key statistic that isn't shown is the frequency at the very last rank, which has been reserved for those tweets for which a recommended hashtag is beyond rank 500. Due to space constraints on the graph, that figure has been omitted. For the graph pertaining to the main file, the frequency is highest at the very last rank. For the Baseline System and the Prior System it is interestingly the same number 6,366.

Figure 2: Main Test File

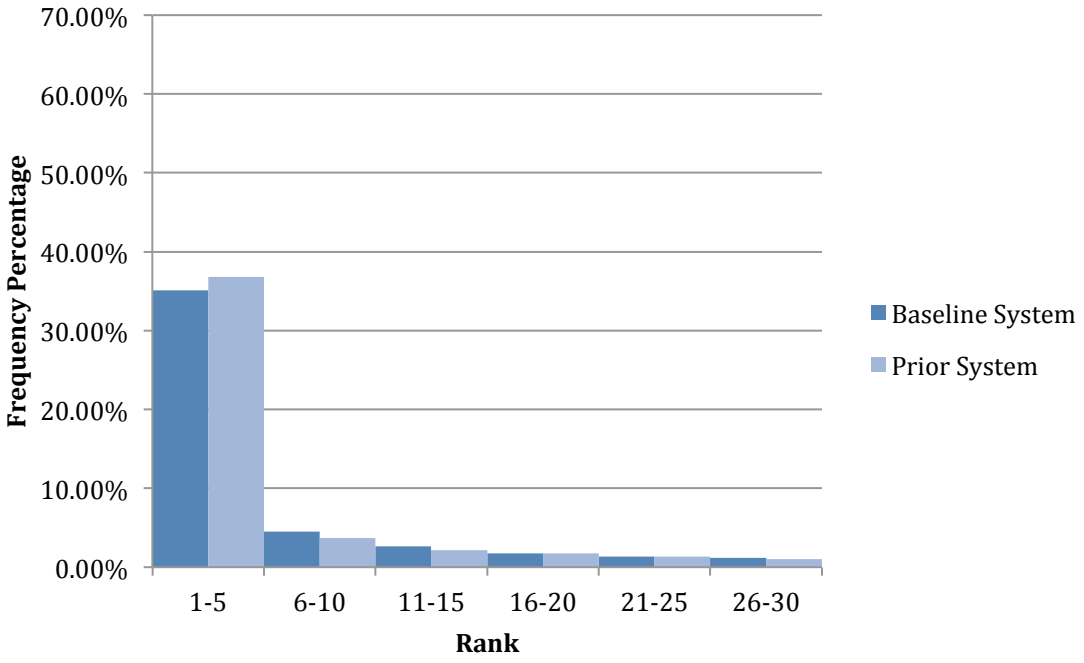


Figure 3: Test file with tokens greater than 9

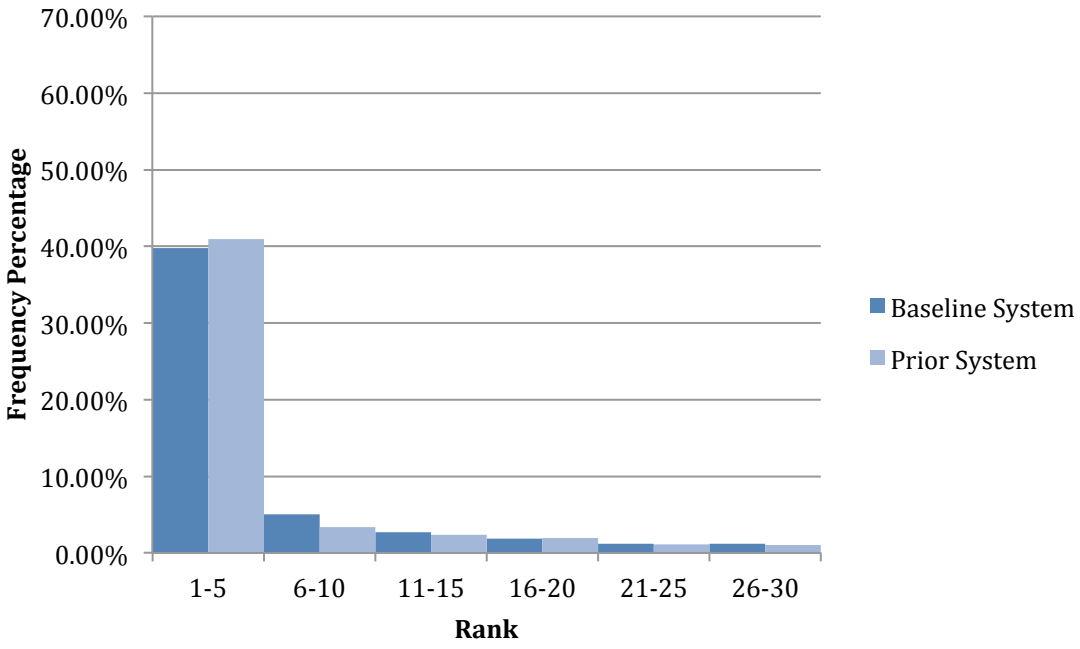


Figure 4: Test File with tokens less than or equal to 9

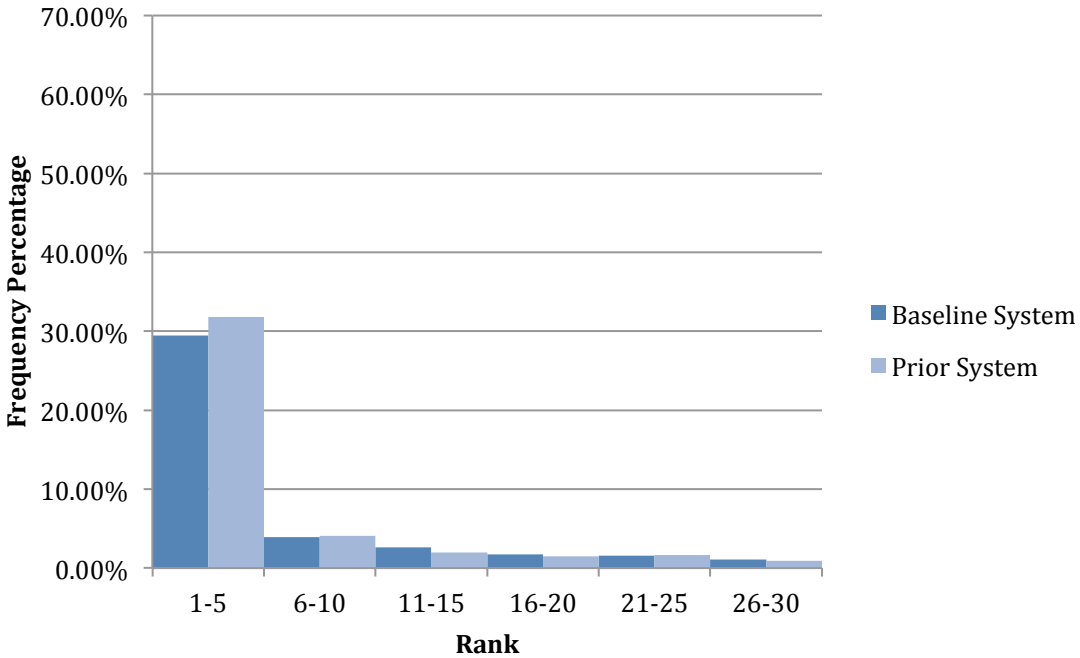


Figure 5: Test File with Single Hastags

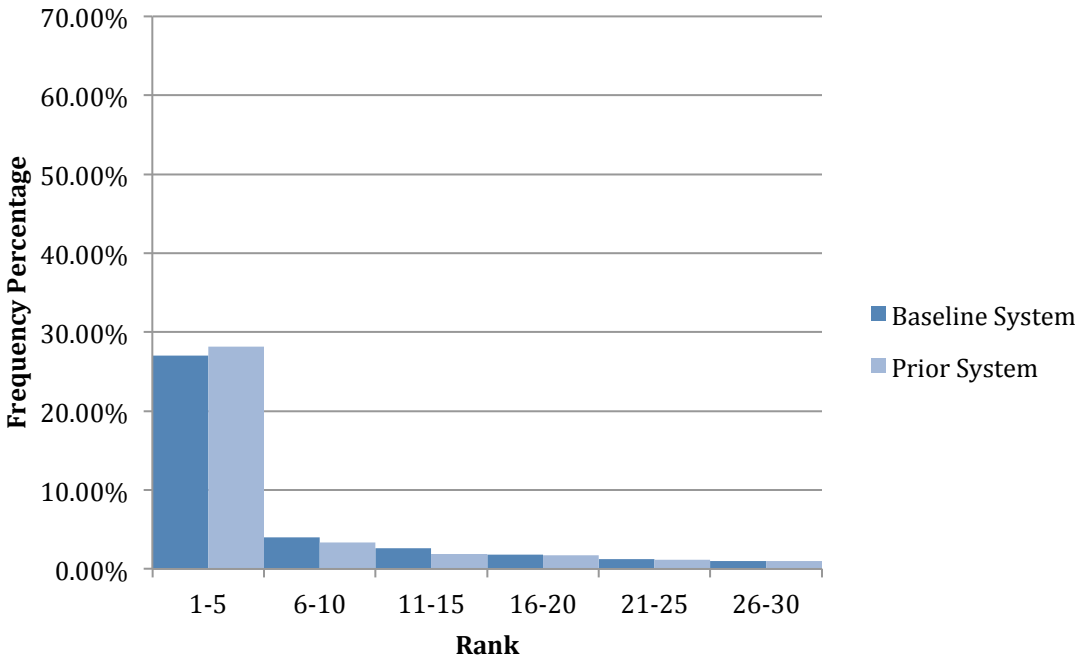


Figure 6: Test File with Multiple Hashtags

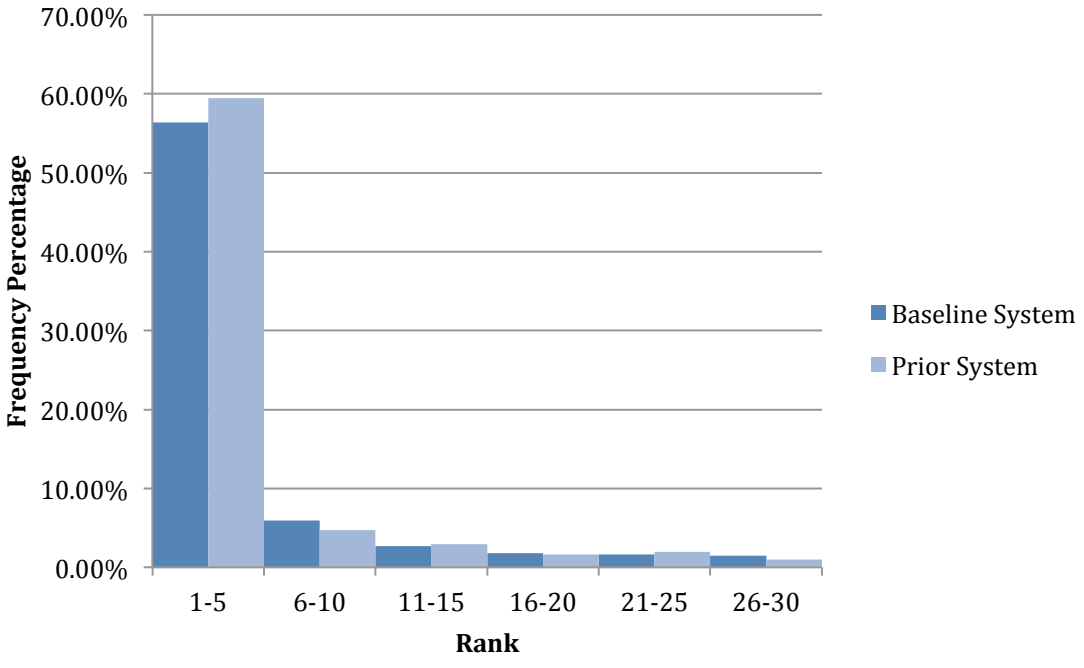


Figure 7: Test File with urls

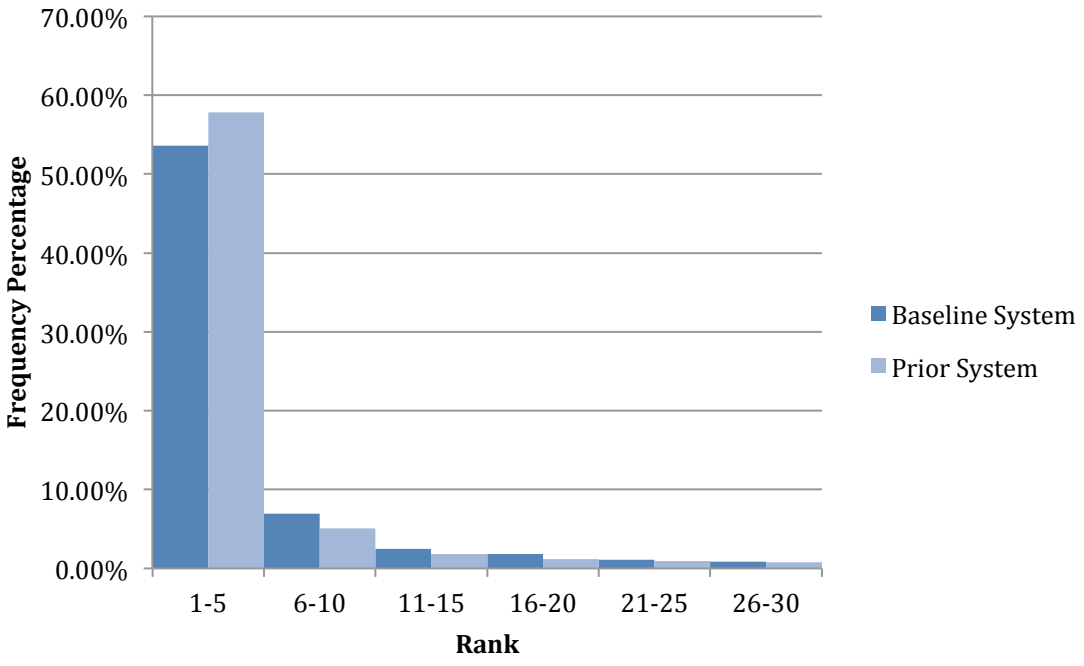


Figure 8: Test Files with no urls

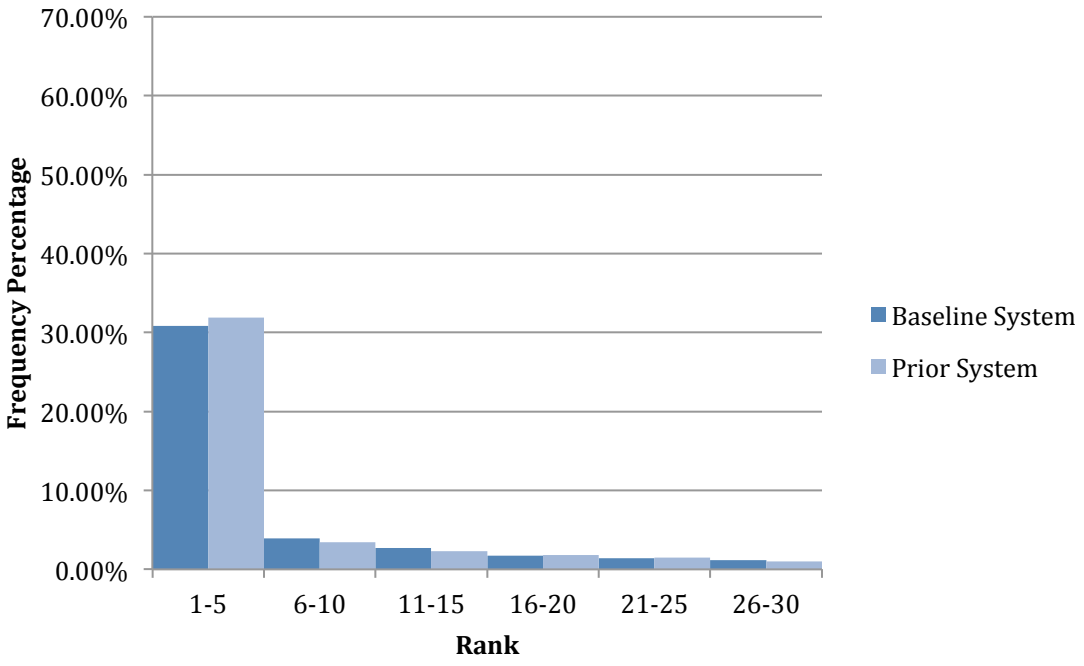


Figure 9: Test File without user mentions

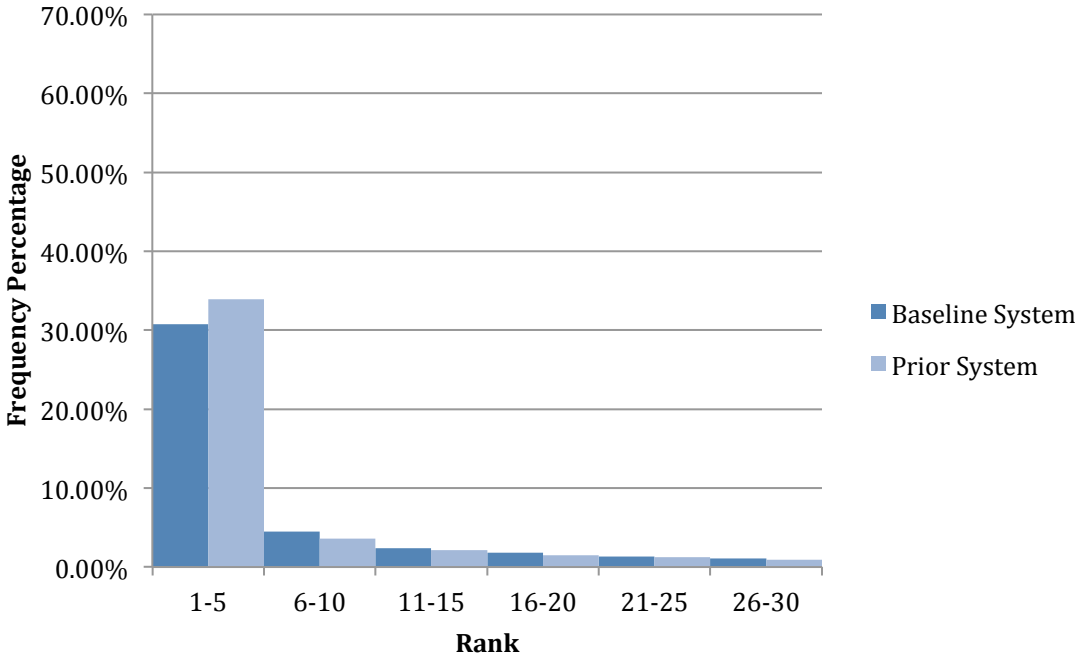
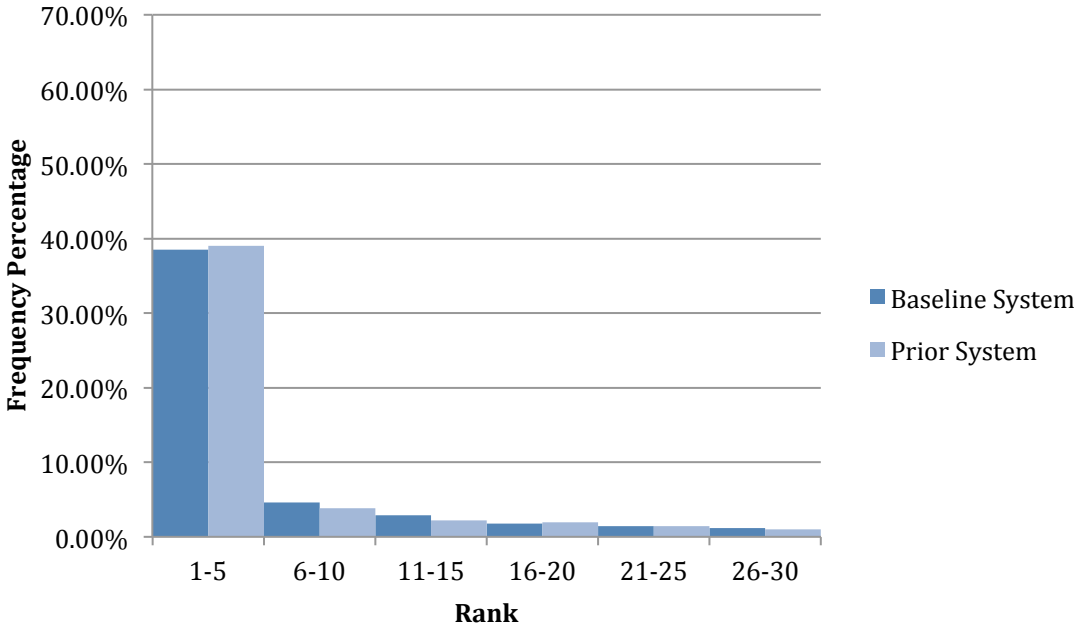


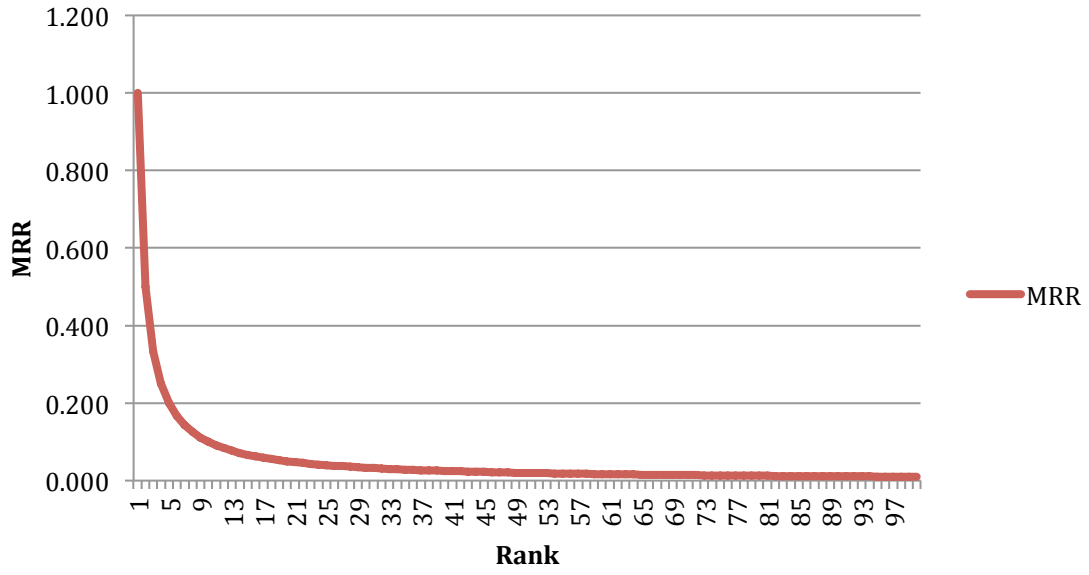
Figure 10: Test File with user mentions



Discussion

From the results above, we can see that there isn't a significant boost to the performance but there was some, and that does show the usefulness of using the hashtag frequency as a prior. As can be observed from the histograms, a significant number of correct predictions were made within the top 5 ranks itself. From an evaluation perspective getting a high percentage of predictions in the top 5 ranks is good as the MRR value starts dropping significantly around rank 5, which can be seen in Figure 11. Another observation is that the Prior System is producing more number of predictions in the top 5 ranks, though by a small margin.

Figure 11: MRR



What is though more interesting is the consistent behavior under both the Baseline System and the Prior System for the cases listed in the problem statement, each of which shall be discussed below.

Tweets with more than 9 tokens vs. tokens less than or equal to 9

This result essentially translates into – More the content in the tweets, easier for the system to draw out the correct hashtag. Since both the system hashtags context are driven by language models, it is not surprising to observe that having more content in the tweet makes it easier for the system to recommend a hashtag better.

Multiple vs. Single Hashtags

From the results we see that having multiple hashtags allows both the algorithms to get a high MRR, this could have been possible due to the fact that out of the multiple hashtags only one was sufficient to have been selected. For example, if a tweet had 3 hashtags, the algorithm would work favorably if 1 out of the 3 came high in the recommendation list.

Having a URL in the tweet vs. not having a URL

Clearly having a URL in the tweet makes it much easier to make a hashtag recommendation. This could most likely be possible as people might use keywords along with the URL to provide some vital information about that URL so that they could pique

people's interest to visit the URL. But this is mere speculation and to closely inspect this further with the numbers we have at our disposal two more tests have been conducted. In the first test, for both the test sets i.e., for one with the URLs and the other without any URLs, the distribution of tokens greater than 9 and less than 9 has been noted. And in the next test, the distribution of hashtag usage i.e., using a single hashtag or multiple hashtags has been noted.

Test File With no URLs	
No. Of tweets with tokens less than or equal to 9	7006
No. Of tweets with tokens more than 9	8463
No. Of tweets with single hashtags	11961
No. Of tweets with multiple hashtags	3508

Test File With URLs	
No. Of tweets with tokens less than or equal to 9	1843
No. Of tweets with tokens more than 9	1672
No. Of tweets with single hashtags	1801
No. Of tweets with multiple hashtags	1714

From both the tables above we can observe that the distribution of the tokens must not be causing the difference in performance due to the fact that they are distributed almost equally. The performance difference could be due to the fact of that multiple hashtags are used more often in comparison to single hashtags when a URL is used in a tweet.

Broadcasted tweets vs. user specific tweets

A possible reason why tweets with user mentions perform better could be that users use much richer content when trying to engage in a community as shown by (Yang, L., et al, 2012), but as with the URLs test sets, this is mere speculation and the distribution of the token usage and hashtag usage needs to be done as in the above comparison to understand this behavior better.

Test File With no User Mentions	
No. Of tweets with tokens less than or equal to 9	4468
No. Of tweets with tokens more than 9	4071
No. Of tweets with single hashtags	6030
No. Of tweets with multiple hashtags	2509

Test File With User Mentions	
No. Of tweets with tokens less than or equal to 9	4405
No. Of tweets with tokens more than 9	6055
No. Of tweets with single hashtags	7747
No. Of tweets with multiple hashtags	2713

Unlike in the previous case with the URLs here the distribution of the token usage is not similar. As can be seen from the table above, more tweets with user mentions seem to be

using more tokens. And as has been already observed, having more tokens boosts the performance of the systems and this could likely be the reason why hashtags having user mentions are performing better.

Conclusion and Future Work

From the experiments the following can be concluded – having more tokens, multiple hashtags, using URLs and mentioning a user in tweets results in the system performing better. A central reason to why tweets with URLs and user mentions perform better is due to the distribution of the tokens and hashtags as seen above. A key aspect that has not been inspected in this study is analyzing tweets that have been completely constructed only with hashtags. It would be interesting to see the impact of one hashtag on the other, and if the positioning of a hashtag does have any impact.

Moving on further, an area I believe that could be analyzed could be the on-topic-ness of a hashtag and the temporal properties of it i.e., how quickly does a particular topic assigned to a hashtag get diluted with content other than the initial topical content, and how soon does another topic replace a current topic related to a hashtag. From an evaluation perspective, I believe a great deal more can be done. The metric that has been used in the experiments only deals with the ability to predict the true hashtag, and provides no other crucial information regarding the hashtags that are surrounding the true hashtag. There definitely might have been cases where the true hashtag, was ranked above 10, but the top 10 hashtags produced were very similar to the true hashtag. In such a scenario, the MRR fails to provide information about such behavior. So I believe a study for a better metric could be done.

References

Zhai, C., & Lafferty, J. (2001, September). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 334-342). ACM.

Efron, M. (2010, July). Hashtag retrieval in a microblogging environment. In*Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 787-788). ACM.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006, August). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia* (pp. 31-40). ACM.

Potts, L., Seitzinger, J., Jones, D., & Harrison, A. (2011, October). Tweeting disaster: hashtag constructions and collisions. In *Proceedings of the 29th ACM international conference on Design of communication* (pp. 235-240). ACM.

Tsur, O., & Rappoport, A. (2012, February). What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In*Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 643-652). ACM.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, *12*(1), 58-72.

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012, April). We know what@ you# tag: does the dual role affect hashtag adoption?. In *Proceedings of the 21st international conference on World Wide Web* (pp. 261-270). ACM.

Guan, Z., Bu, J., Mei, Q., Chen, C., & Wang, C. (2009, July). Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 540-547). ACM.

Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., & Van de Walle, R. (2013, May). Using topic models for Twitter hashtag recommendation. In*Proceedings of the 22nd international conference on World Wide Web companion* (pp. 593-596). International World Wide Web Conferences Steering Committee.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM*, *10*, 10-17.

MacArthur, A. The History of Hashtags. Retrieved from: http://twitter.about.com/od/Twitter-Hashtags/a/The-History-Of-Hashtags.htm

Van Grove, J. (2011, January). Instagram Introduces Hashtags for Users & Brands. Retrieved from: http://mashable.com/2011/01/27/instagram-hashtags/

Di Caro, L., Candan, K. S., & Sapino, M. L. (2008, August). Using tagflake for condensing navigable tag hierarchies from tag clouds. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1069-1072). ACM.

Quintarelli, E. (2005). Folksonomies: power to the people.

Pirolli, P. (2005). Rational analyses of information foraging on the web.*Cognitive science*, *29*(3), 343-373.

Spero, S. (2013). My Boston Marathon Experience Through Twitter. Retrieved from: http://blog.havasdiscovery.com/index.php/my-boston-marathon-expericene-through-twitter/

Kywe, S. M., Hoang, T. A., Lim, E. P., & Zhu, F. (2012). On recommending hashtags in twitter networks. In *Social Informatics* (pp. 337-350). Springer Berlin Heidelberg.

Zangerle, E., Gassler, W., & Specht, G. (2011). Recommending#-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings* (Vol. 730, pp. 67-78).

Ding, Z., Zhang, Q., & Huang, X. (2012). Automatic Hashtag Recommendation for Microblogs using Topic-Specific Translation Model. In *COLING (Posters)*(pp. 265-274).

Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., & Van de Walle, R. (2013, May). Using topic models for Twitter hashtag recommendation. In*Proceedings of the 22nd international conference on World Wide Web companion* (pp. 593-596). International World Wide Web Conferences Steering Committee.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., ... & Riedl, J. (2006, November). Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 181-190). ACM.

Farhi, P. (2009). The Twitter Explosion-Whether they are reporting about it, finding sources on it or urging viewers, listeners and readers to follow them on it, journalists just

can't seem to get enough of the social networking site. Just how effective is it as a journalism tool?. *American Journalism Review (AJR)*,*31*(3), 26.

Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010, December). Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips* (Vol. 104, No. 45, pp. 17599-601).

Hong, L., Dan, O., & Davison, B. D. (2011, March). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web* (pp. 57-58). ACM.

Sen, S., Vig, J., & Riedl, J. (2009, April). Tagommenders: connecting users to items through tags. In *Proceedings of the 18th international conference on World wide web* (pp. 671-680). ACM.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006, January). Folkrank: A ranking algorithm for folksonomies. In K. D. Althoff (Ed.), *LWA* (Vol. 1, pp. 111-114).

Mishne, G. (2006, May). Autotag: a collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the 15th international conference on World Wide Web* (pp. 953-954). ACM.